Machine Learning 1

Lecture 7.2 - Supervised Learning
Classification - Probabilistic Discriminative
Models - Logistic Regression

Erik Bekkers

(Bishop 4.3.2)

Slide credits: Rianne van den Berg

Image credit: Kirillm | Getty Images

Classification Strategies

- *Discriminant functions*

Direct mapping of input to target $t = y(\underline{x}, \underline{w})$

- *Probabilistic discriminative models*

Posterior class probabilities: $p(C_k \mid \underline{x})$

- *Probabilistic generative models* $p(\underline{x}, C_k)$

Class-conditional densities: $p(\underline{x} \mid C_k)$

Prior class probabilities: $p(C_k)$

$\left.\begin{array}{l} p(\underline{x} \mid C_k) \\ p(C_k) \end{array}\right\} \overset{\text{Bayes}}{\Longrightarrow} p(C_k \mid \underline{x})$

Logistic Regression for Two Classes

- Given: Dataset $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_N)^T$ with binary targets

  $\mathbf{t} = (t_1, ..., t_N)^T$ with $t_n \in \{\mathcal{C}_1, \mathcal{C}_2\} = \{1, 0\}$

- Basis functions $\boldsymbol{\phi} = \boldsymbol{\phi}(\mathbf{x}) = (\phi_0(\underline{x}), ..., \phi_{M-1}(\underline{x}))^T$

  $\phi_0(\mathbf{x}) = 1$

- Probabilistic Discriminative Linear Models: posteriors $p(\mathcal{C}_k|\boldsymbol{\phi})$ are nonlinear functions with a linear function of $\boldsymbol{\phi}$ as input.

$$p(\mathcal{C}_k|\boldsymbol{\phi}, \mathbf{w}) = f(\underbrace{\mathbf{w}^T\boldsymbol{\phi}}_{\text{linear}})$$

non-linear

Logistic sigmoid

Generative setting

$p(C_1|\underline{x}) = \sigma\left(\ln \frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)}\right)$

- Logistic regression (K=2)

  $C = C_1$

$$p(\mathcal{C}_1|\boldsymbol{\phi}, \mathbf{w}) = y(\boldsymbol{\phi}) = \sigma(\mathbf{w}^T\boldsymbol{\phi})$$

  $C = C_2$

$$p(\mathcal{C}_2|\boldsymbol{\phi}, \mathbf{w}) = 1 - y(\underline{\phi}) = 1 - \sigma(\underline{w}^T\underline{\phi})$$

$$p(t|\boldsymbol{\phi}, \mathbf{w}) = y(\phi)^t(1 - y(\phi))^{1-t}$$

Logistic Regression for Two Classes

$\mathbf{w}, \boldsymbol{\phi} \in \mathbb{R}^M$

- Logistic Regression:

$$p(\mathcal{C}_1|\boldsymbol{\phi}, \mathbf{w}) = \sigma(\mathbf{w}^T\boldsymbol{\phi}) \qquad p(\mathcal{C}_2|\boldsymbol{\phi}, \mathbf{w}) = 1 - \sigma(\mathbf{w}^T\boldsymbol{\phi})$$

parameters: $\mathbf{w}: M$ linearly with M

- Gaussian conditional densities:

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}_k|^{1/2}}\exp\{\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\}$$

class priors $p(\mathcal{C}_k)$

parameters: $\boldsymbol{\mu}_k : M$

$\Sigma : \sim M^2$ quadratically with M

Logistic Regression for Two Classes

- Given: Dataset $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_N)^T$ with binary targets
  $\mathbf{t} = (t_1, ..., t_N)^T$ with $t_n \in \{\mathcal{C}_1, \mathcal{C}_2\} = \{1, 0\}$

- Conditional likelihood function:

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}) = \prod_{n=1}^{N} p(t_n | x_n, \underline{w}) = \prod_{n=1}^{N} y_n^{t_n} (1-y_n)^{1-t_n}$$

$$y_n = p(\mathcal{C}_1|\boldsymbol{\phi}_n) = \sigma(\underline{w}^T \underline{\phi_n}) \qquad \boldsymbol{\phi}_n = \boldsymbol{\phi}(\mathbf{x}_n)$$

- Maximizing the conditional likelihood/minimizing the cross-entropy

(Bishop 4.3.3)

(Newton > 0)

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}) = -\sum_{n=1}^{N} t_n \ln y_n + (1-t_n) \ln(1-y_n)$$

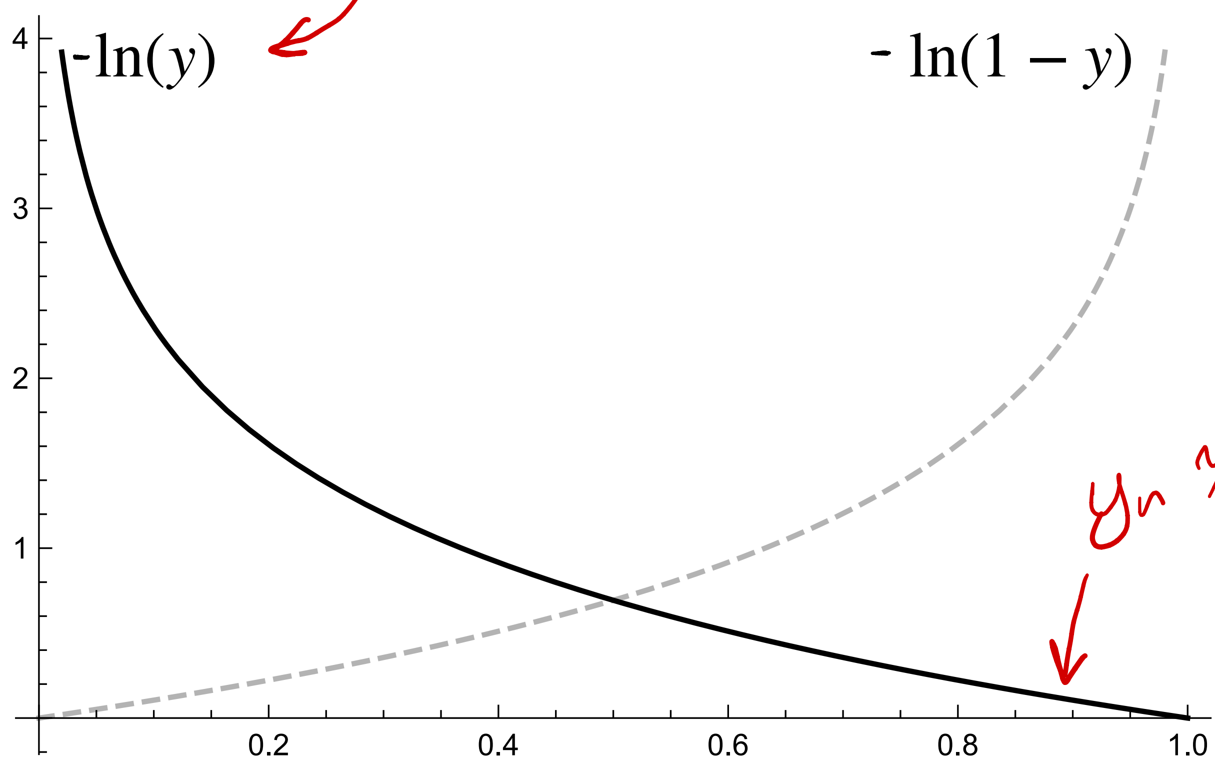
- E(**w**): convex, but no closed form solution!

$$y_n = \sigma(\mathbf{w}^T \boldsymbol{\phi}_n) \text{ is nonlinear in } \mathbf{w}$$

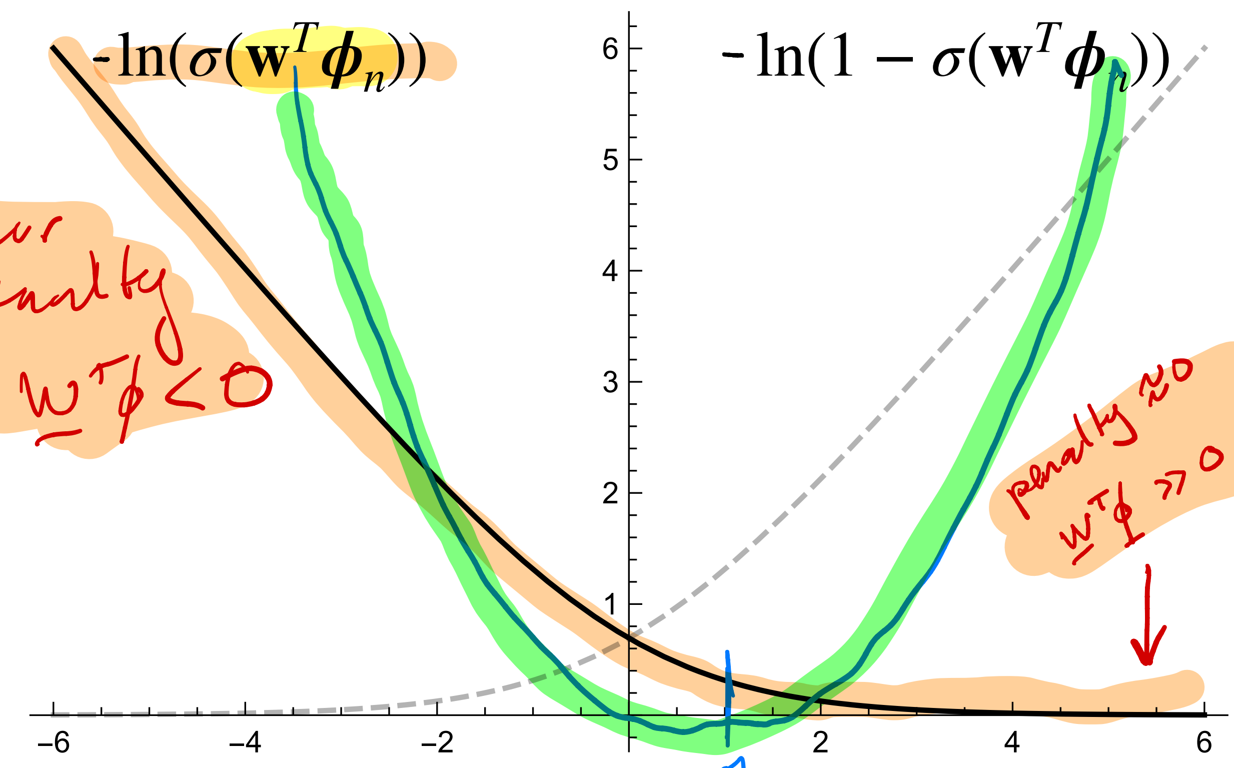
The cross-entropy loss

$$E_n(\mathbf{w}) = -\left\{t_n \ln(y_n) + (1 - t_n)\ln(1 - y_n)\right\}$$

target $t=1$



$t=1$



LS regression

$y = w^T\phi$

loss

$(y-t)^2$

Least squares outlier problem does not take place with logistic regression

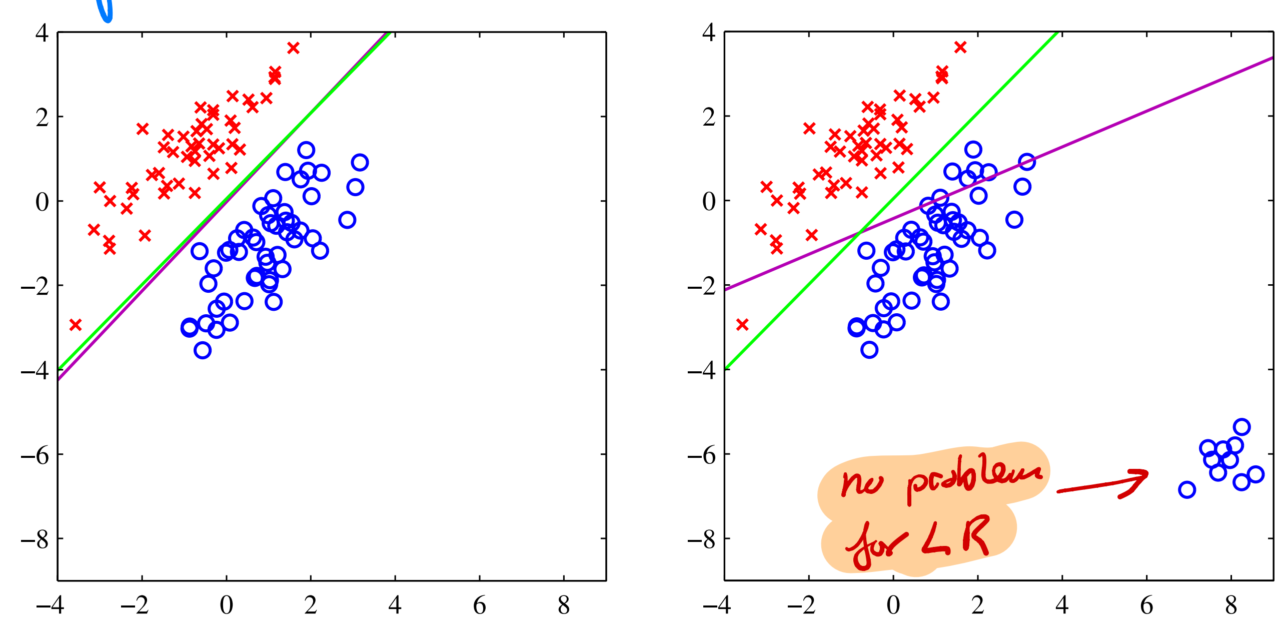


Figure: least squares is very sensitive to outliers (Bishop 4.4)

Classification with Logistic Regression

- Dataset: $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_N)^T$ with targets
  $\mathbf{t} = (t_1, ..., t_N)^T$ with $t_n \in \{\mathcal{C}_1, \mathcal{C}_2\} = \{1, 0\}$
- Basis functions $\boldsymbol{\phi}(\mathbf{x}) = (\phi_0(\mathbf{x}), ..., \phi_{M-1}(\mathbf{x}))^T$
- Posterior distributions: $p(\mathcal{C}_1|\mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}))$
- Minimize $E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}) = -\sum_{n=1}^{N} t_n \ln y_n + (1 - t_n) \ln(1 - y_n)$
  with stochastic gradient descent or iterative reweighted least squares, to find $\mathbf{w}^*$ (next videos)
- New datapoint $\mathbf{x}'$ is assigned to class C_1 if
  $p(\mathcal{C}_1|\mathbf{x}', \mathbf{w}^*) = \sigma((\mathbf{w}^*)^T \boldsymbol{\phi}(\mathbf{x}')) > \frac{1}{2}$ $\quad \underline{w}^{*T} \underline{\phi}(\underline{x}) > 0$
- Decision boundaries:
  $\underline{w}^{*T} \underline{\phi}(\underline{x}) = 0$